International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

APPLICATIONS AND CLASSIFICATION RULES OF GENETIC ALGORITHM IN DATA MINING

M SIVAMANI
Research Scholar
Department of Computer Science
Sri Vasavi College, Erode
Smksivamani@gmail.com

ABSTRACT- Data mining has as goal to extract knowledge from large databases. To extract this knowledge, a database may be considered as a large search space, and a mining algorithm as a search strategy. In general, a search space consists of an enormous number of elements, making an exhaustive search infeasible. Therefore, efficient search strategies are of vital importance. Search strategies based on genetic based algorithms have been applied successfully in a wide range of applications. In this paper, we discuss the suitability of genetic-based algorithms for data mining. We discuss the various application areas genetic Algorithm plays evolutionary role with data mining technique and explain them in details.

Keywords: Genetic algorithm, Classifier, Data mining, Classification Genetic Algorithms

INTRODUCTION

The Genetic Algorithm was developed by John Holland in 1970.There are no known polynomial time algorithms to solve many real-world optimization problems making them hard to solve a number of heuristics have been designed to solve the hard problems. These heuristics may provide sub optimal but acceptable solution in a reasonable computational time. A number of meta-heuristics such as simulated annealing, evolutionary algorithms, artificial networks derived from natural physical and bio logical phenomena have also been used to solve these problems

DATA MINING

A field that deals with extracting knowledge from databases, without putting restrictions on the amount or types of data in a database, is data mining. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful.

DATA: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.

These includes operational or transactional data such as, sales, cost, inventory, payroll, and

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

accounting nonoperational data, such as industry sales, forecast data, and macro economic data
Meta data - data about the data itself, such as logical database design or data dictionary definitions

INFORMATION: The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

KNOWLEDGE: Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

DATA WAREHOUSES: Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

CLASSIFICATION: Classification is a form of Data Analysis that can be used to construct a Model, which can be further used in future to predict the Class Label of new Datasets. Various Application of classification includes Fraud Detection, Target Marketing, Performance Prediction, Manufacturing and Medical Diagnosis. Data Classification is a two step process

(i) The first step is a learning step. In this step a classification algorithm builds the Classifier by Analyzing (or learning from) a training set made up of database tuples and their associated Class Labels. In this first step a Mapping Function Y=f(X) is learned that can predict the associated Class Label Y of a given tuple X. That mapping function or Classifier can be in the form of Classification Rules, Decision Trees or Mathematical Formulae.

(ii) Next Step of Classification, Accuracy of a Classifier is predicted. For this another set of tuples

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

apart from training tuples are taken called as Test Sets. Then these set of tuples of test set are given as

Input to the Classifier. The Accuracy of a Classifier on a given test set is the percentage of test set Tuples that are correctly classified by the Classifier.

DATA MINING WORK

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships such as

Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

APPLICATION OF GENETIC ALGORITHM

Genetic Algorithms is an effective tool to use in data mining and pattern recognition. There are two different methods to applying Genetic algorithm in pattern recognition.

1 Use Genetic algorithm as a classifier directly in computation.

2. Use a Genetic algorithm to optimize the results i.e. as an optimizer to arrange the parameters in other classifiers. Most applications of GAs in pattern recognition optimize some parameters in the classification process Genetic algorithms has been applied to find an optimal set of feature weights that improve classification accuracy. First, a traditional feature extraction method such as Principal Component Analysis (PCA) is applied, and then a classifier such as k-NN (Nearest Neighbor Algorithm) is used to calculate the fitness

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

function for GA [10], [7]. Combination of classifiers is another area that GAs have been used to optimize. GA is also used in selecting the prototypes in the case-based classification.

According to us second method of genetic algorithm to optimize the result from the dataset is more effective to compute the accurate values of observations of data by applying data mining techniques.
Genetic Algorithm has a wide scope in business

There are large amount of data that has to be filtered to process the results for optimizing the business profits by using various data mining techniques. There are many domains in business to which they can be applied:

I. Optimization: Give a business problem with certain variables and a well defined definition of profit, a genetic algorithm can be used to automatically determine the optimal value for the variables that optimize the profit

II. Prediction: Genetic algorithms have been used as Meta level operators that are used to help optimize other data mining algorithms. For instance, they have been used to find the optimal association rules in market-analysis

III. Simulation: Sometimes a specific business problem is not well defined in terms what the profit is or whether one solution is better than the other. The business person instead just has large number of entities that they would like to simulate via simple interaction rules overtime.

a) Classification Methods in Data Mining

Classification by Decision Tree Induction: In this method Decision Tree is learned from Class Labeled training tuples and then it is used for Classification. A Decision Tree is a flowchart–like tree structure, where each Internal Node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. While learning the Decision Tree or we can say during Tree Construction, attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes. Once Decision Trees are built, Tree Pruning attempts to identify and remove branches that may reflect noise or outliers in the training data.

Rule Based Classification: Rule Based Classifiers uses a set of IF-Then Rules for Classification. IF Condition Then Conclusion. The IF part is the Rule Antecedent or Pre condition. Then Part is the Rule Consequent. The Condition consists of one or more Attribute Tests that are logically AND. The Rule's

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

Consequent contains a Class Prediction. Let D be the Training data set. Let X is a tuple .If a Rule is satisfied by X, the rule is said to be triggered .Rule fires by returning the Class Prediction. Rule Extraction from a Decision Tree to Extract Rules from a Decision Tree, One Rule is created for each path from the root to a leaf node. Each Splitting Criterion along a given path is Logically AND to form the Rule Antecedent (IF part).The Leaf node holds the Class Prediction, forming the Rule Consequent (Then part).Rule Induction using a Sequential Covering Algorithm Here the Rules are Learned Sequentially, One at a time (for one Class at a time) directly from the Training Data (i.e. without having to generate a Decision Tree first) using a Sequential Covering Algorithm.

CONCLUSION AND FUTURE SCOPE

The first and most important point is that genetic algorithms are intrinsically parallel. Most other algorithms are serial and can only explore the solution space to a problem in one direction at a time since GAs have multiple offspring, they can explore the solution space in multiple directions at once. If one path turns out to be a dead end, they can easily eliminate it and continue work on more promising avenues, giving them a greater chance each run of finding the optimal solution. Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. It has been shown to be efficient and powerful through many data mining applications that use optimization and classification. GAs can rapidly locate good solutions, in data mining even for difficult search spaces. GAs are used in various fields of Data mining to get the optimized solutions for the better performance of the data that are required in decision making and process the accurate result. There is also a greater scope of GA in data mining in future application to stimulate the data mining concepts. Genetic algorithms are widely applicable to classification by means of inductive learning. GAs also provides a practical method for optimization of data preparation and data transformation steps. Hence GA can be used in a real analysis system to get the better solution.

The comprehensibility of the discovered patterns (features) could be improved with a proper modification of the fitness function. How much predictive accuracy would be increased with such a modification is a question whose answer requires further research.

Future work should consist of more experiments with other data sets, as well as more elaborated experiments to optimize several parameters of the algorithm, such as mutation rates, the Limit threshold for the weight field, etc.

REFERENCES

1. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., 1996.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 4, 2017 ISSN: 2348-6600
Reference ID: IJCS-174 PAGE NO: 1063-1068
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

Advances in Knowledge Discovery and Data Mining. MIT Press.

2. Forrest, Stephanie. "Genetic algorithms: principles of natural selection applied to computation." Science, vol.261, p.872-878 (1993).

3. Pei, M., Punch, W.F., and Goodman, E.D. "Feature Extraction Using Genetic Algorithms", Proceeding of International Symposium on Intelligent Data Engineering and Learning'98 (IDEAL'98), Hong Kong, Oct. (1998).